

How much raw sounds can do for word learning: learning by ‘repeating’ in a human-robot interaction

Yo Sato, Ze Ji, Hagen Lehmann and Chrystopher L. Nehaniv

I. BACKGROUND

Extracting linguistic units from raw speech sound is an essential part of language acquisition, and its importance in its own right—as dissociated from meaning in particular—is increasingly recognised [1]. Such purely phonological sound form learning nevertheless has not yet attracted much attention in the study of learning agents, including developmental robotics. Furthermore, most attempts in the field of machine learning that purport to model infants’ phonological word discovery have hitherto presuppose *phonemes* to be readily available ([2], [3], [4]), but as the long-standing difficulty to phonemise sounds in speech processing research demonstrates, phonemes are linguistic abstractions, specific to a language and not mechanically recoverable, and hence needs to be learnt. The present work is an attempt to address the question of how words may be learnt without recourse to anything else than what you find in the sound data, not even phonemes.

Recently such language-independent methods for speech sound learning have started to emerge ([5], [6]). In the study described in what follows we use two representative aspects amongst them: acoustics and statistics. We further add to these the element of *interaction* using human-robot interaction (HRI), in response to the suggestion [7] that the trials and errors by actually producing the sound units the learner predicts to be valid completes the learning loop.

Our model posits three stages for learning of sound forms that the learner predicts to be a linguistic unit—word-like units for short. At the first, *holistic perception* stage, the learner identifies and stores in memory recurrent signal patterns found in raw speech data. Being ‘holistic’ this learning does not result in any linguistic unit but simply extracts fragments of acoustic stream which correspond to *holophrases*. Discovery of linguistic units are left to the second, *constructive representation* stage, where the learner becomes sensitive to the segmental constituents and the statistical regularities thereof, and builds up expectations about the make-up of a *syllable*. The learner now becomes ready for the third, *production and interaction* stage, where it actually tries out the phone concatenations in the syllabic form to the caregiver and adjusts their expectations according to the response from him/her.

This paper presents an implementation of this model that culminates in an HRI experiment currently conducted, after we describe the mechanisms for the first two stages it presupposes.

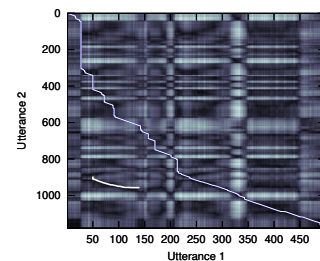
II. HOLISTIC PERCEPTION: HOLOPHRASE DETECTION

We do not only assume that the only resource pre-linguistic children initially have access to is raw acoustic input but

also avoid attempting to extract phonemes directly from there, firstly practically, to avoid the same well-known technical difficulty in phonemising sounds, and secondly to reflect the observation that the mastery in phonemes indeed come much later in children’s speech development [8]. A more simple and domain-general approach is to detect *repetitions* in speech. Along with [9] and [10], we use the Dynamic Time Warping (DTW) to search for similar acoustic patterns. Essentially, we slice up each utterance into tiny temporal segments, from which to extract the normalised acoustic features (mel-frequency cepstrum), and calculate the feature distance (the shorter the ‘closer’) for the segments for a *pair* of utterances. We conduct the comparisons for all the possible pairs in a set of utterances in search for a *valley* as a holophrase candidate, a continuous segments where the distances are small.

While some researchers associate the found recurrent patterns with cognitively higher-level processing (such as with referents [6]), we stick to the acoustic level and explore the effectiveness of extracting word-like units by invoking *prosodic* information. That is, we hypothesise that the learner focuses its attention to the

loudest part—usually a vowel—of an utterance and start the search for a valley from that point. The search goes bidirectionally, as it is not predictable where in the potential valley the prominent point is located. The figure above shows the found valley in white line, which is /læks/ in ‘A black square, can you see’. We have thus far successfully extracted a number of keywords in the pilot study. These matched streams are stored in memory for the use in the subsequent stages.



III. CONSTRUCTIVE REPRESENTATION: BREAKING DOWN AND RECONSTRUCTING SOUNDS

This stage concerns the process of building mental representations of sound units, so that the learner will develop the readiness to reproduce what it has heard. Here the learner, having isolated and stored recurrent streams, now tries to ‘imitate’ them, but importantly, the learner cannot yet ‘copy and reproduce’ the desired sound sequence since it cannot yet reliably predict what sounds will be produced by its own articulation. The learner needs first to abstract out from the acoustic data the representations it uses for production.

To model the perceptual situation, we use the narrow phonetic transcription without training to extract phonetic segments. To simulate the abstraction of linguistic representations then, we apply a classification learning based on cluster analysis to arrive at groupings of consonants and vowels and obtain the phone-segmented data.¹

At this point we would be ready to let the learner utter the found recurrent matches using its phonetic constituents, but we take an intermediate step: *syllables*. This is partly due to the plausible view that they form the basis for speech generation [13], but more importantly, because they make children’s utterances *productive*. It is well attested that infants do not always imitate but often utters simple syllables not corresponding to any words or phrases yet compliant with the phonotactics of the language they are exposed to ([14], [15]), revealing the *constructive* nature of children’s babbling.

For syllable discovery, we invoke *transitional probability* between phones, which Saffran et al. [16] famously showed infants are sensitive to. The learner then uses this statistic to induce the likely boundaries between phones, building thereby the expectation as to what constitutes a syllable, or ‘mental syllabary’ in the sense described in [13].

We test two scenarios under which the learner utters a word-like unit: *holophrastic repetition* and *syllabic repetition*. The first manner takes place when the learner detects a repetition in the preceding utterance of a signal pattern that was perceived in the past, as described in the preceding section. Now this does not always happen, as it depends on the frequency that the participants repeat a word or phrase, on their articulateness and so on. In the absence of a recurrent match, we invoke the syllabic repetition method, in which the learner simply repeats the most stressed ‘syllable’ in the immediately preceding utterance. In either case, the learner produces an extracted ‘syllable’ — a probable C*VC* sequence (C denotes a consonant, V vowel, and * zero or more occurrences)— for a required number of times (once for the latter case, and as many times as syllables for the former case).

IV. PRODUCTION AND INTERACTION: FEEDBACK EFFECTS

We have incorporated into a humanoid robot the mechanisms described above for acoustic recurrent holophrase detection and finding representational units, and started a series of HRI experiments. The objective of the experiments is to simulate the caregiver-infant scenario, so as to evaluate the effects of feedbacks from the caregiver. In the HRI sessions that use the iCub robot [17], designed to have similar features to a human infant, each participant is asked to ‘teach’ it 3-D shapes using foam blocks. We selected shapes the words for which are composed of one to three syllables, namely *star*, *ball*, *donut*, *prism*, *cylinder* and *pyramid*.

The robot waits for the end of an utterance in the participant’s speech, and then responds by ‘reproducing’ the prosodically salient part of the utterance that has been just made

¹We currently follow the following procedure: use HTK [11] for phonetic transcription; turn it into feature-based representations of phones; apply the K-nearest neighbor algorithm on them; and use the phonetic-capable TTS system (Festival, [12]) for production. Admittedly this does not faithfully reflect the cognitive process we plan to switch to an articulatory speech synthesiser.

by the participant, in two ways described in the preceding section. If a recurrence has been detected, it produces (the phonetic reconstruction of) the matched holophrase, while if not, it produces (the reconstruction of) the salient syllable in the immediately preceding utterance. Thus the dialogue may go like this:

- | | | |
|--------------|---------------------------------|-----|
| Participant: | This is a cýlinder, Deechee. | |
| iCub: | [ɪ] | (1) |
| Participant: | Cylinder, Deechee, cýl-lin-der. | |
| iCub: | [[ɪ]] | (2) |
| Participant: | Good trý, Deechee, cylinder. | |
| iCub: | [səlëndə] | (3) |
| Participant: | Well dóne Deechee. | |
| iCub: | [daŋ] | (4) |

where “ ’ ” marks the most prominent vowel. In (1) and (2), the iCub fails to find a match in a previous utterance, so utters (a reconstruction of) the most stressed syllable of the preceding utterance. In (3), it does find one, thanks to the repetition of a keyword, *cylinder*. What we pay particular attention to is the effects of *corrective feedback*, which would indirectly provide the learner with negative evidence.

REFERENCES

- [1] P. Jusczyk, *Discovery of spoken language*. MIT Press, 1997.
- [2] M. Brent and T. Cartwright, “Distributional regularity and phonotactic constraints are useful for segmentation,” *Cognition*, vol. 61, pp. 93–125, 1996.
- [3] C. Yu and D. Ballard, “A multimodal learning interface for grounding spoken language in sensorimotor experience,” *ACM Transactions on Applied Perception*, vol. 1, pp. 57–80, 2004.
- [4] S. Goldwater, T. Griffiths, and M. Johnson, “A Bayesian framework for word segmentation: Exploring the effects of context,” *Cognition*, vol. 112, pp. 21–54, 2009.
- [5] M. Huckvale, I. Howard, and S. Fagel, “Klair: a virtual infant for spoken language acquisition research,” in *Proceedings 10th Interspeech Conference*, 2009.
- [6] L. ten Bosch, H. V. hamme, L. Boves, and R. K. Moore, “A computational model of language acquisition: the emergence of wordsfundamenta informaticae pp. 229249, 2009.” *Fundamenta Informaticae*, vol. 90, 2009.
- [7] M. Vihman, “Word learning and the origins of phonological systems,” in *Language Acquisition*, S. Foster-Cohen, Ed. MacMillan, 2009.
- [8] V. Hazan and S. Barrett, “The development of phonemic categorization in children aged 6-12,” *Journal of phonetics*, vol. 24, pp. 377–396, 2000.
- [9] A. S. Park and J. R. Glass, “Unsupervised pattern discovery in speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.
- [10] G. Aimetti, L. t. Bosch, and R. K. Moore, “The emergence of words: Modelling early language acquisition with a dynamic systems perspective,” in *10th Annual Conference of the International Speech Communication Association – Interspeech*, Brighton, UK, 2009.
- [11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for Version 3.4)*. Cambridge University Department of Engineering, 2004.
- [12] R. A. Clark, K. Richmond, and S. King, “Festival 2 – build your own general purpose unit selection speech synthesiser,” in *Proc. 5th ISCA workshop on speech synthesis*, 2004.
- [13] W. Levelt and L. Wheeldon, “Do speakers have access to a mental syllabary,” *Cognition*, 1994.
- [14] D. K. Oller, L. Wieman, W. Doyle, and C. Ross., “Infant babbling and speech,” *Journal of Child Language*, vol. 3, pp. 1–11, 1976.
- [15] B. de Boysson-Bardies, *How language comes to children*. MIT Press, 1999.
- [16] J. Saffran, E. Aslin, and R. Newport, “Word segmentation: The role of distributional cues,” *Journal of Memory and Language*, vol. 35, 1996.
- [17] G. Sandini, G. Metta, and D. Vernon, “The icub cognitive humanoid robot: An open-system research platform for enactive cognition,” in *50 Years of AI*. Springer, 2007.